



# ***Data – Who Needs It?***

**Keynote presentation for the 2006  
Data Intensive Computing Environment  
Vendor Day**

**March 6, 2006**

**Bill Kramer**

**NERSC Center General Manager, LBNL**





# The HPC Community Must Address Three Trends

- The widening **gap** between application **performance** and peak performance of high-end computing systems
- The recent emergence of **large, multidisciplinary** computational science teams in the DOE research community
- The **flood of** scientific **data** from both simulations and experiments, and the convergence of computational simulation with experimental data collection and analysis in complex workflows

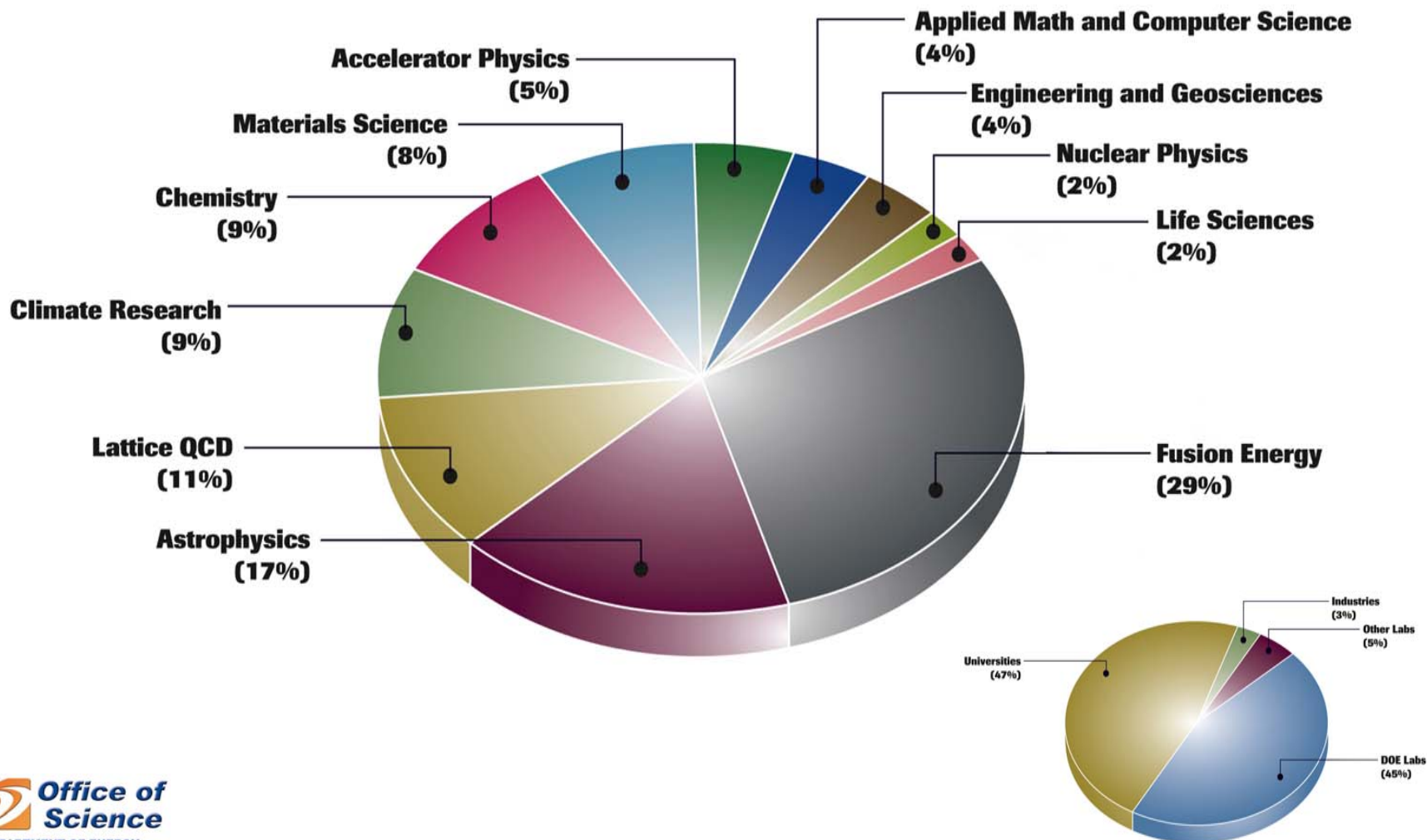


# NERSC Science-Driven Computing Strategy 2006 -2010





# NERSC Usage by Discipline (2005)





# What is Analytics and Why Do We Need It?

- **Analytics is the intersection of:**
  - Visualization, analysis, scientific data management, human-computer interfaces, cognitive science, statistical analysis, reasoning, ...
- **All sciences need to find, access, and store and understand information**
- **In some sciences, the data management (and analysis) challenge already exceeds the compute-power challenge in required resources**
- **The ability to tame a tidal wave of information will distinguish the most successful scientific, commercial, and national security endeavors**
- **It is the limiting or the enabling factor for a wide range of sciences**



# NERSC's Analytics Strategy

- **Broad strategic program objectives:**
  - Clear picture of user needs
  - Leverage existing and provide new visualization and analysis capabilities
  - Enhance data management infrastructure
  - Enhance distributed computing infrastructure
  - Realizing analytics: support for the computational science community

## Composition of the Cosmos



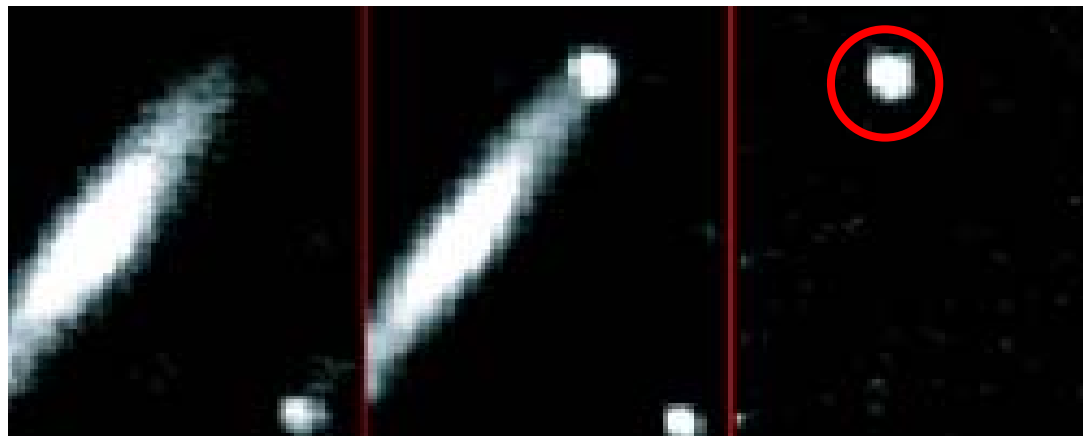


# Distributed Analytics Workflow

## Example: SNFactory

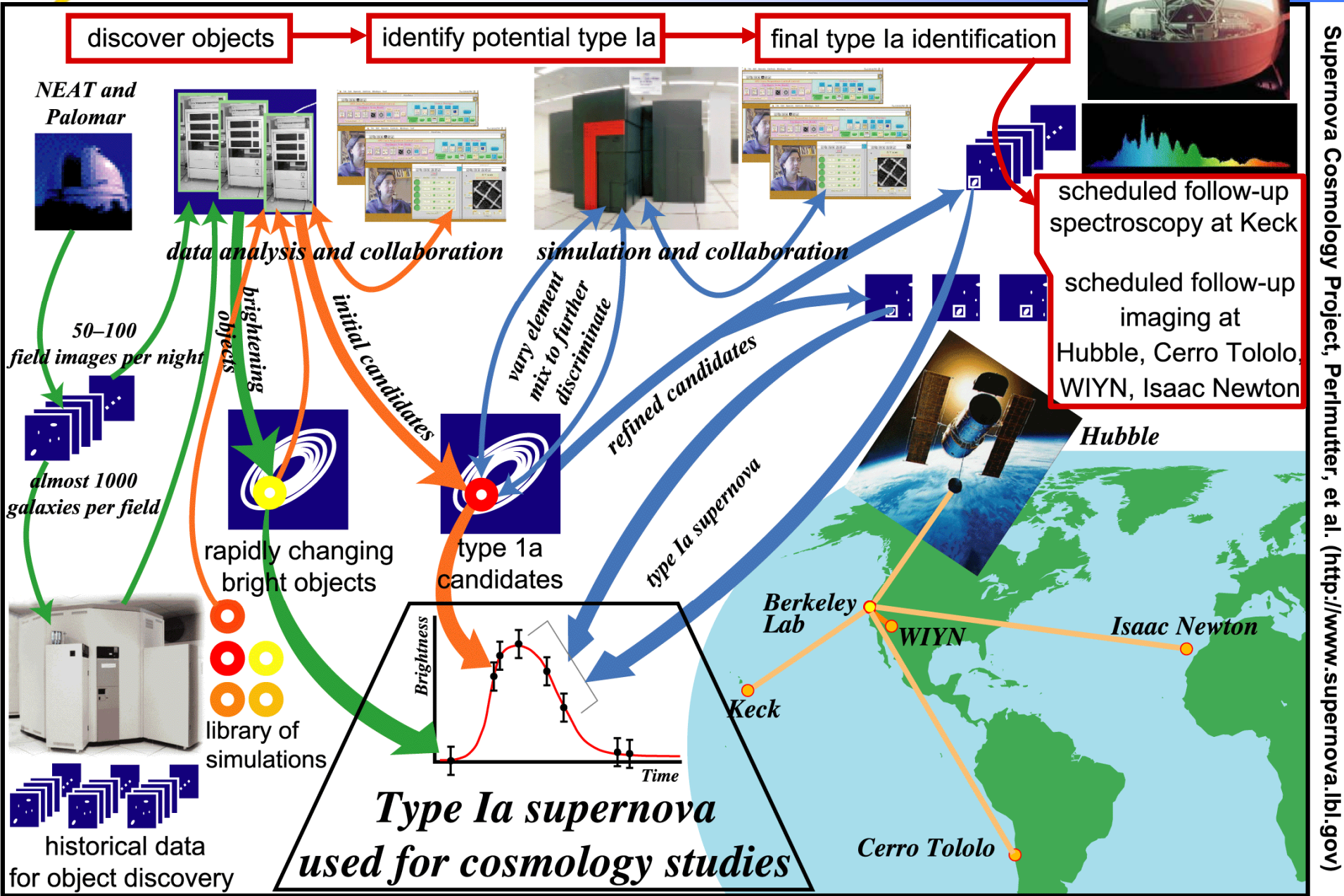
*Distributed Analytics Workflow serves an entire community*

- Images collected from NEAT (Near- Earth Asteroid Tracking) telescopes
- First year: processed 250,000 images, archived
  - 6 TB of compressed data!
- Images sent from telescope to network via custom wireless network
- Images sent to NERSC for analysis on PDSF. Digital processing (registration, differencing) to locate potential targets
- Potential Type 1a supernovae targets identified and broadcast to observation community (24-hour turnaround)





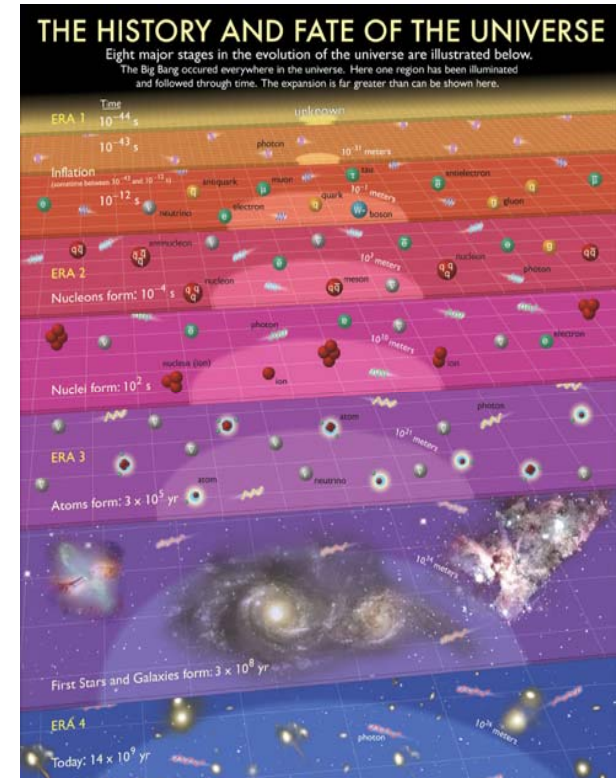
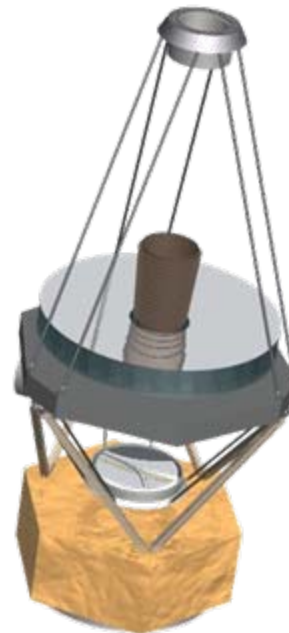
# Supernova Cosmology Depends on Cyberinfrastructure





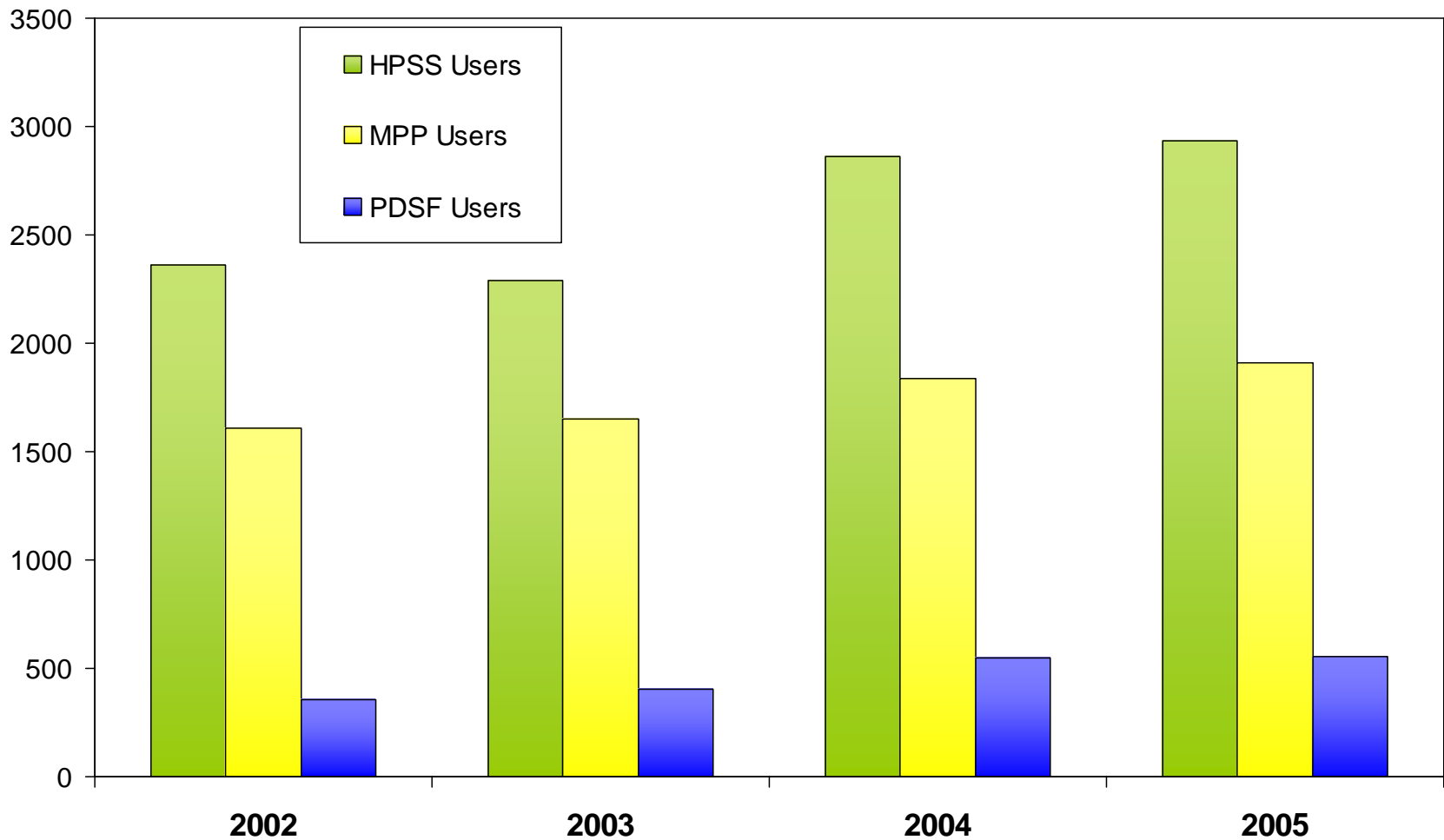
# Simulation Science by 2010: Cosmic Simulator

- Science driven vision of a computational framework in 2010.
  - The Cosmic Simulator is the concept of providing an integrated framework in which component simulations can be linked together to provide a coherent, end-to-end history of the Cosmos.
- SuperNova Acceleration Probe
  - A satellite planned for flight in 2009
- Planck CMB Mission
  - Satellite Mission planned for 2008 will product 10,000,000 pixels and require  $10^{21}$  flop/s for processing.





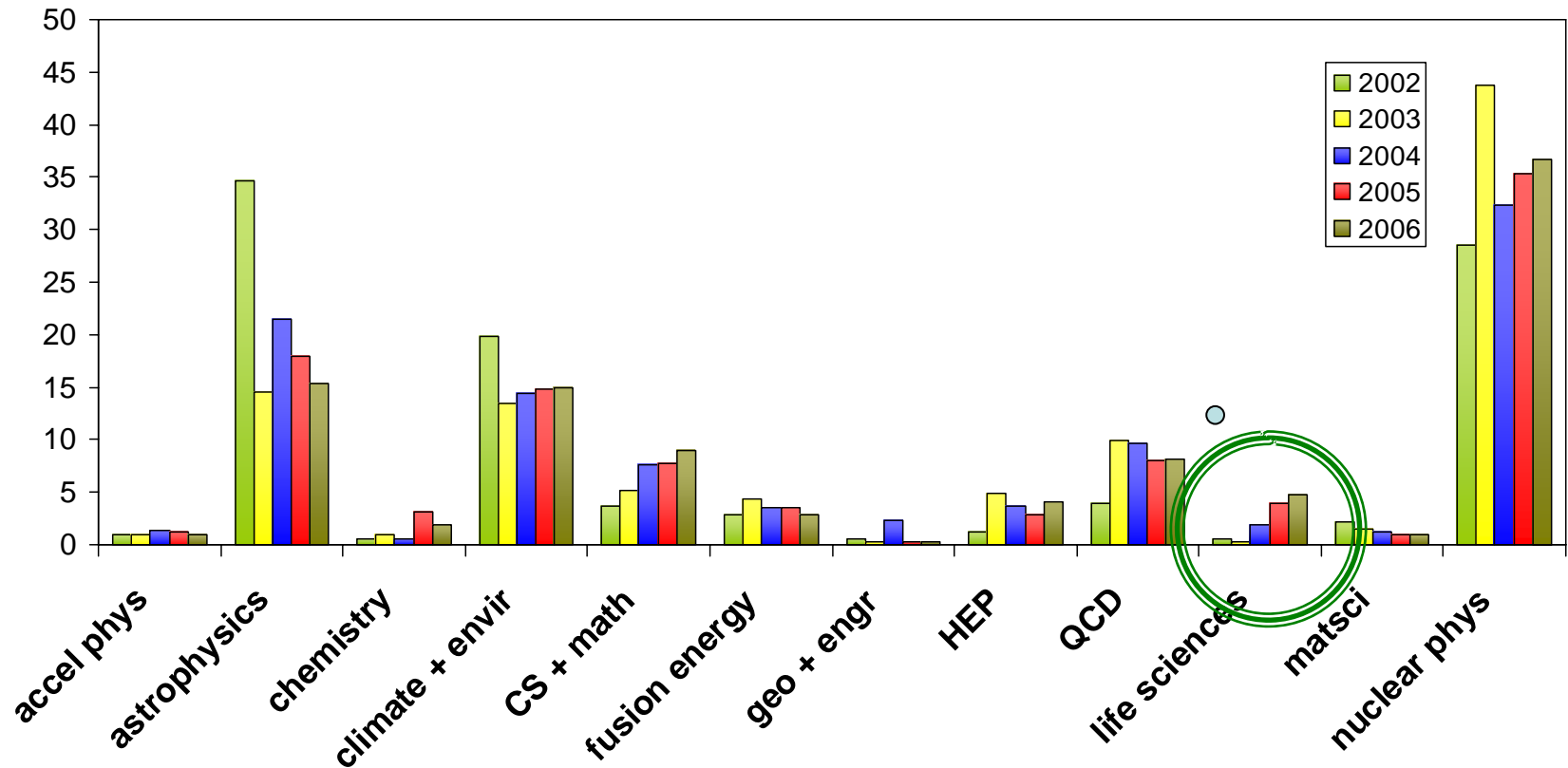
# Increasing Numbers of NERSC Storage Users





# Storage Use

Percent of HPSS Allocation to Science Areas



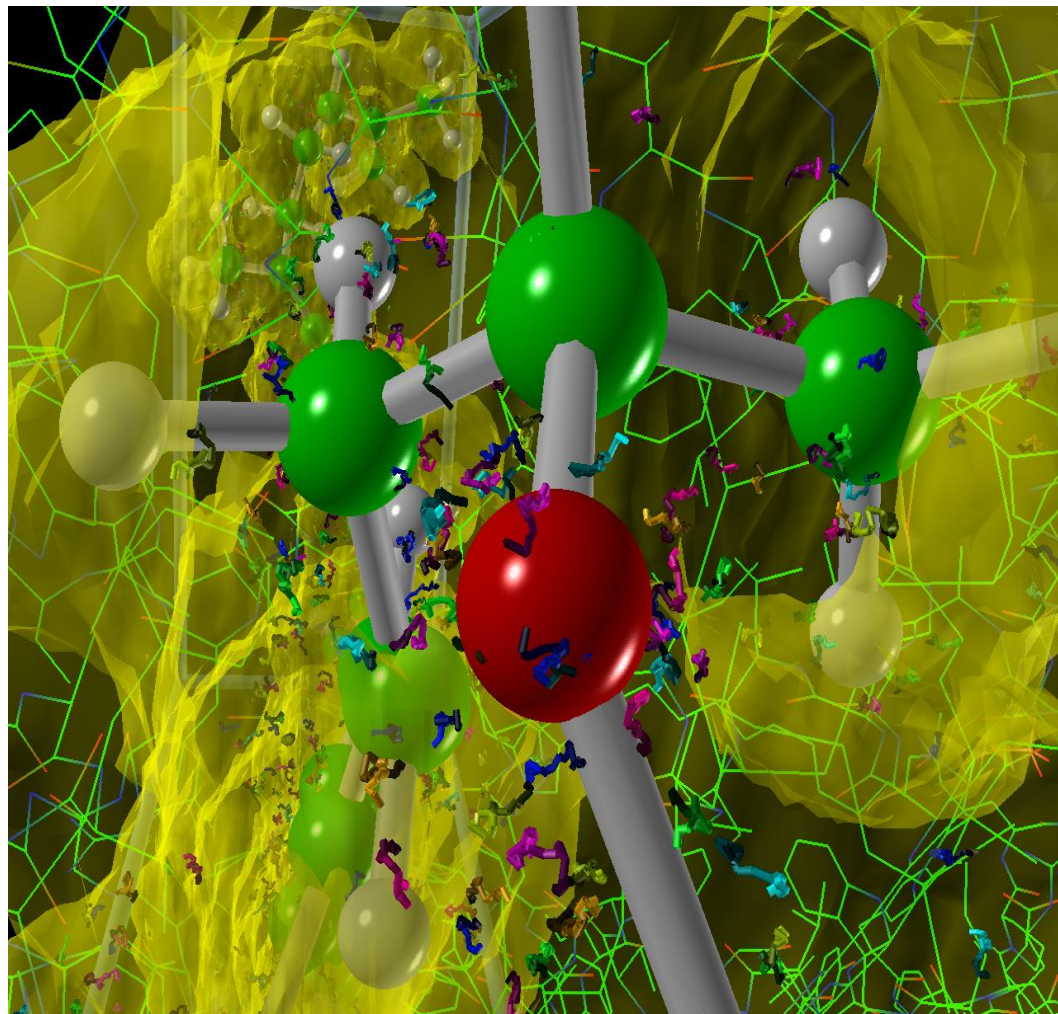


# Photosynthesis INCITE Project

- MPI tuning: 15-40% less MPI time
- Quantum Monte Carlo scaling: 256 to 4,096 processors
- More efficient random walk procedure
- Wrote parallel HDF layer
- Used AVS/Express to visualize molecules and electron trajectories
- Animations of the trajectories showed 3D behavior of walkers for the first time

*“Visualization has provided us with modes of presenting our work beyond our wildest imagination”*

*“We have benefited enormously from the support of NERSC staff”*



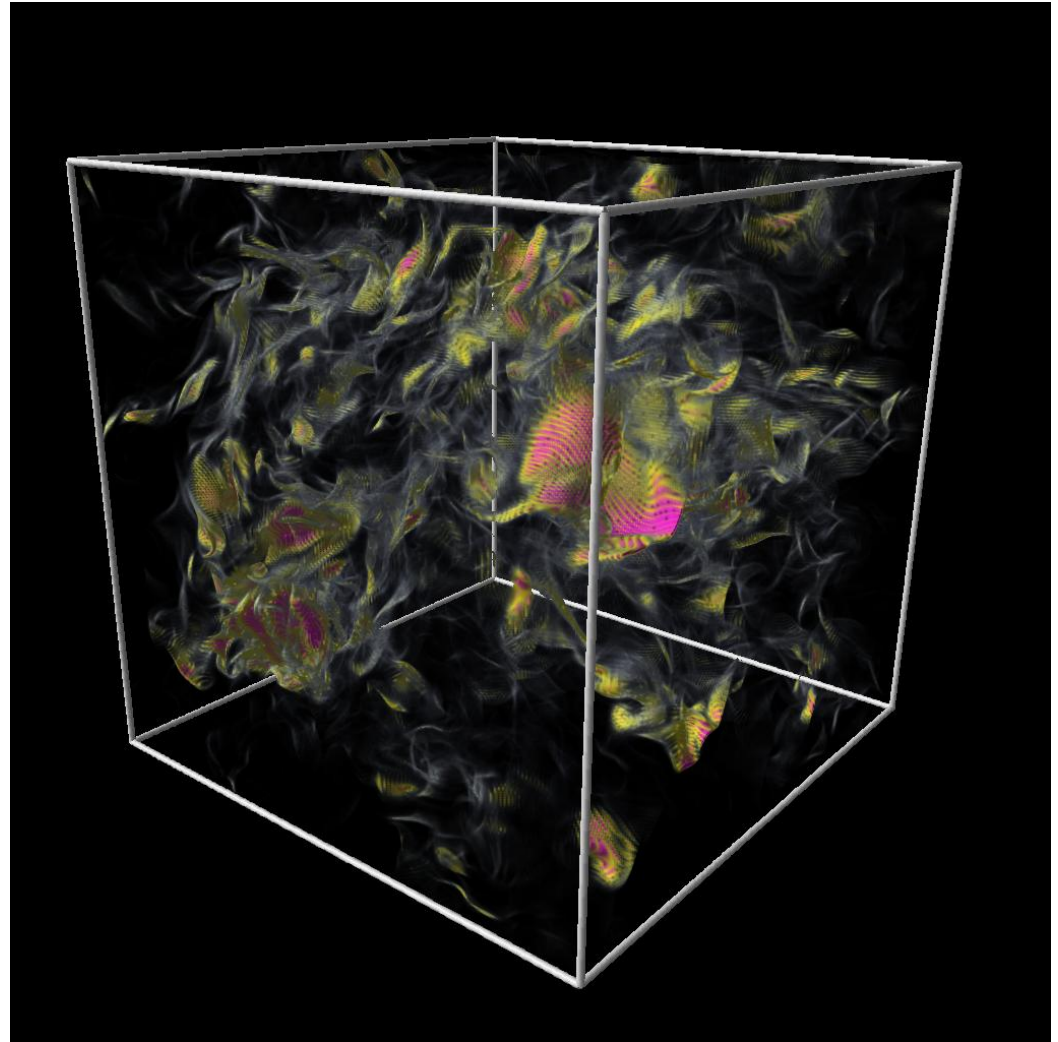


# Fluid Turbulence INCITE Project

- Reduced memory requirements and added threaded FFT: allowed group to solve larger and more interesting problems
- Visualization challenge: simulations produce large and feature-rich time-varying 3D data
- Vis solution: use Ensign parallel backend and Ensign client locally - collaboration resulted in deployment of Remote Vis License server

*“We really appreciate the priority privilege that has been granted to us in job scheduling”*

*“The consultant services are wonderful. We have benefited from consultants’ comments on code performance, innovative ideas for improvement, and diagnostic assistance”*

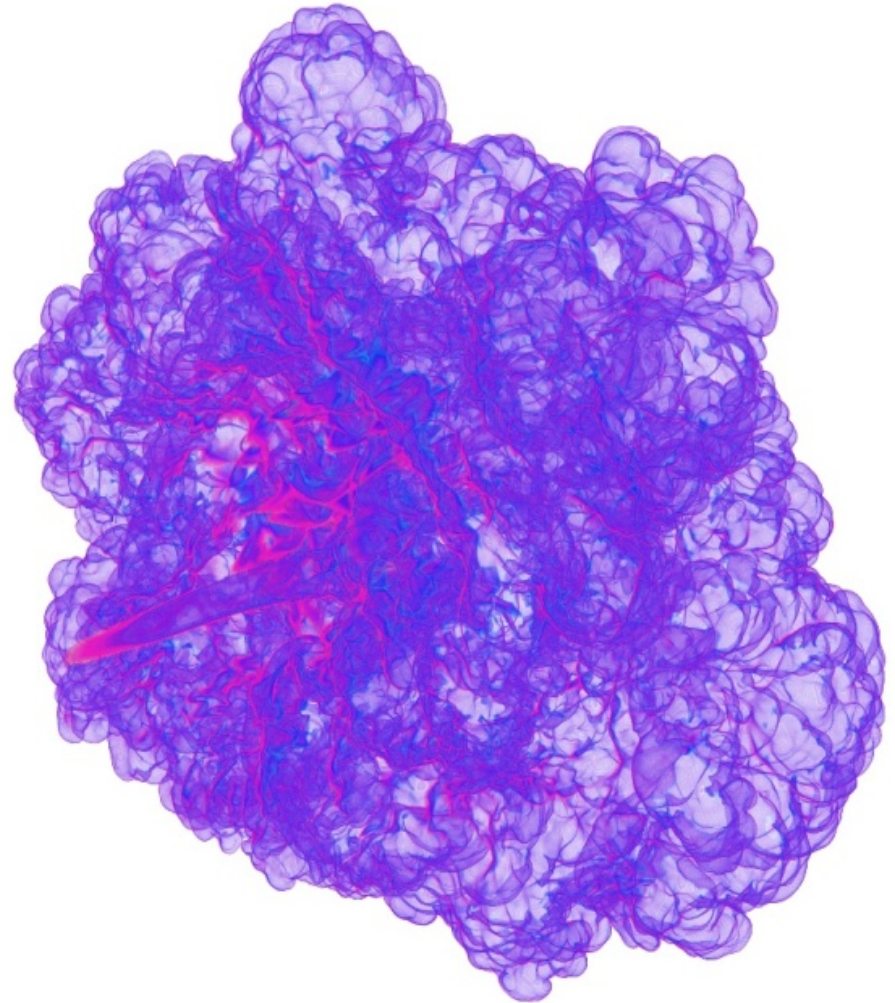




# Thermonuclear Supernovae INCITE Project

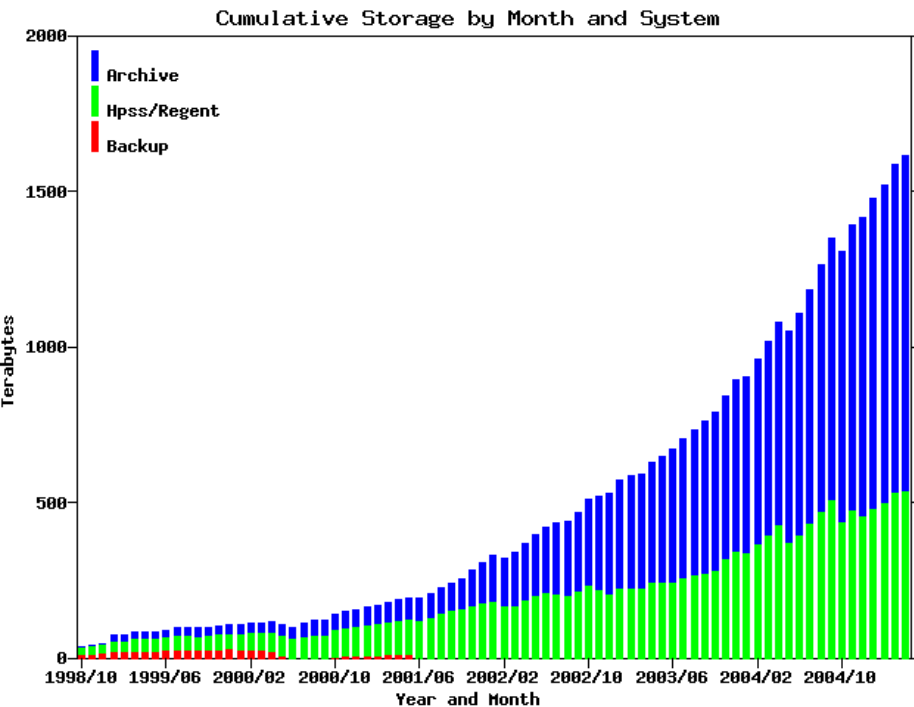
- Resolved problems with large I/O by switching to a 64-bit environment
- Tuned network connections for major storage: transfer rate went from 0.5 to 70 MB/sec
- Created automatic procedure for code checkpointing

*“We have found NERSC staff extremely helpful in setting up the computational environment, conducting calculations, and also improving our software”*



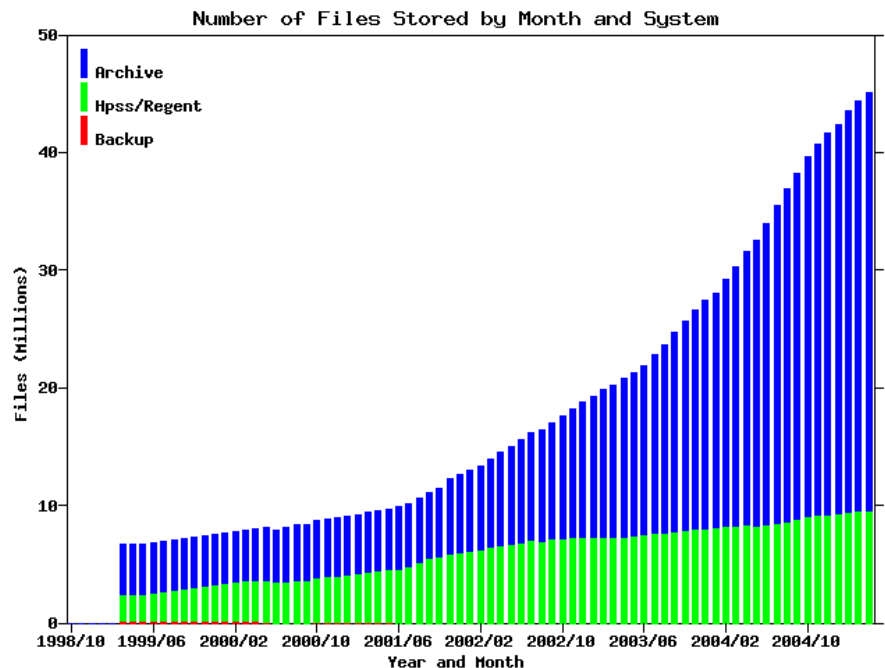


# NERSC Archive Storage Growth



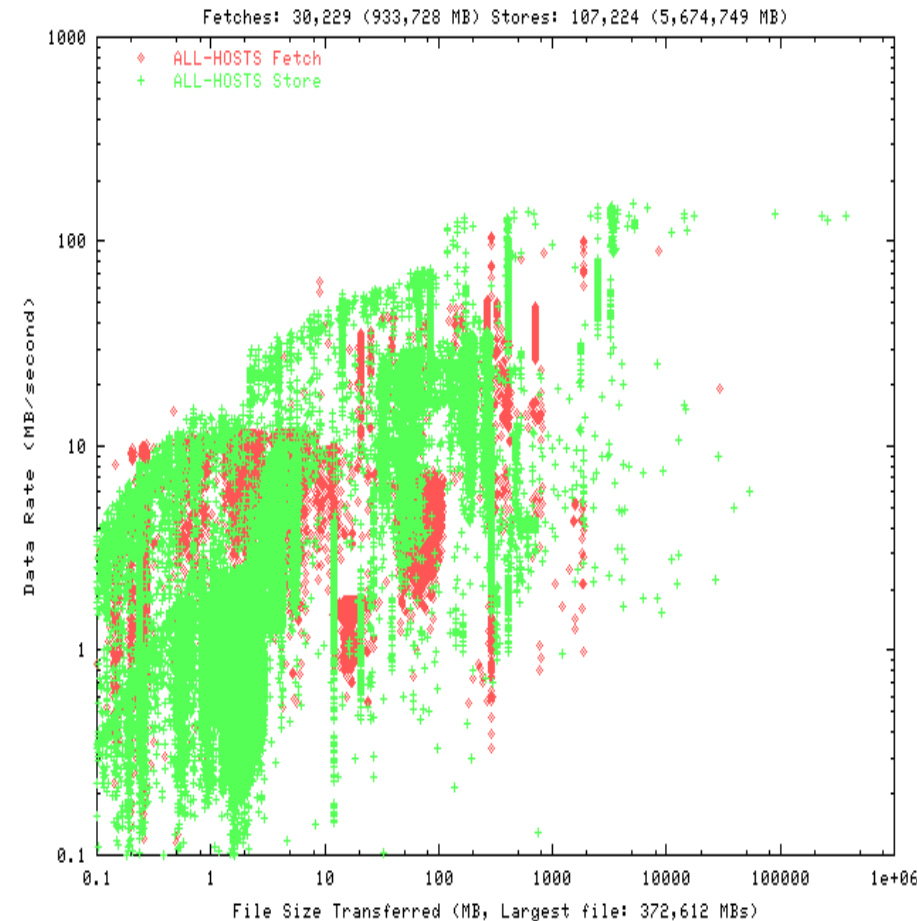
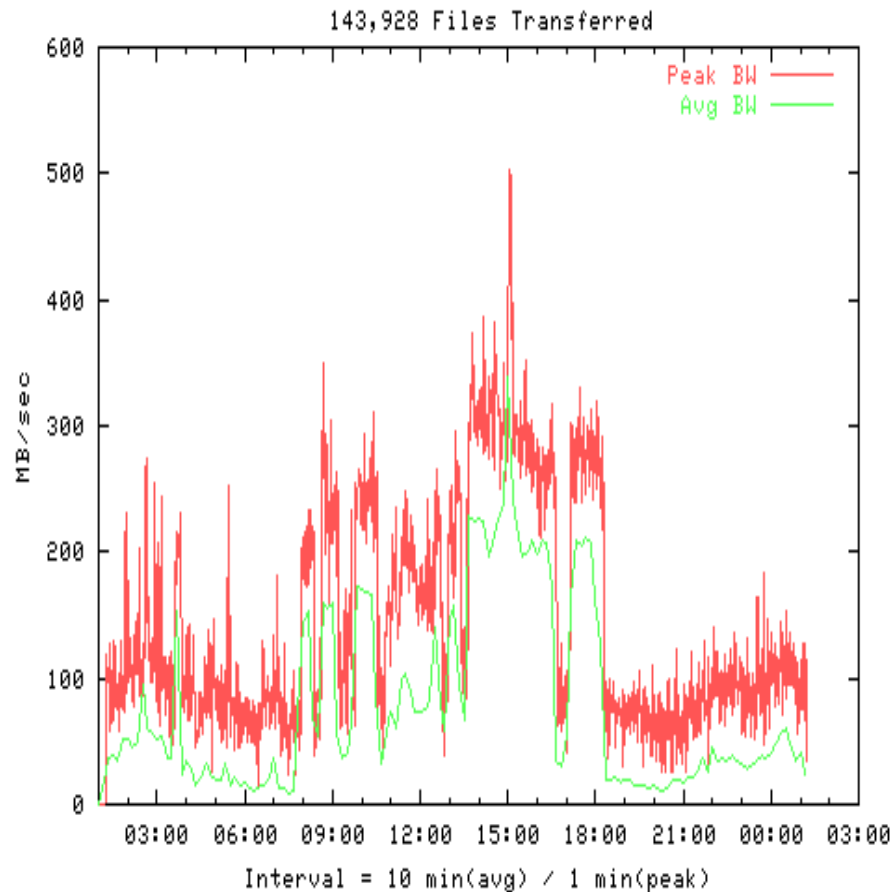
Increase:  
1.7x per year in data growth

45 million files makes NERSC  
one of the largest sites





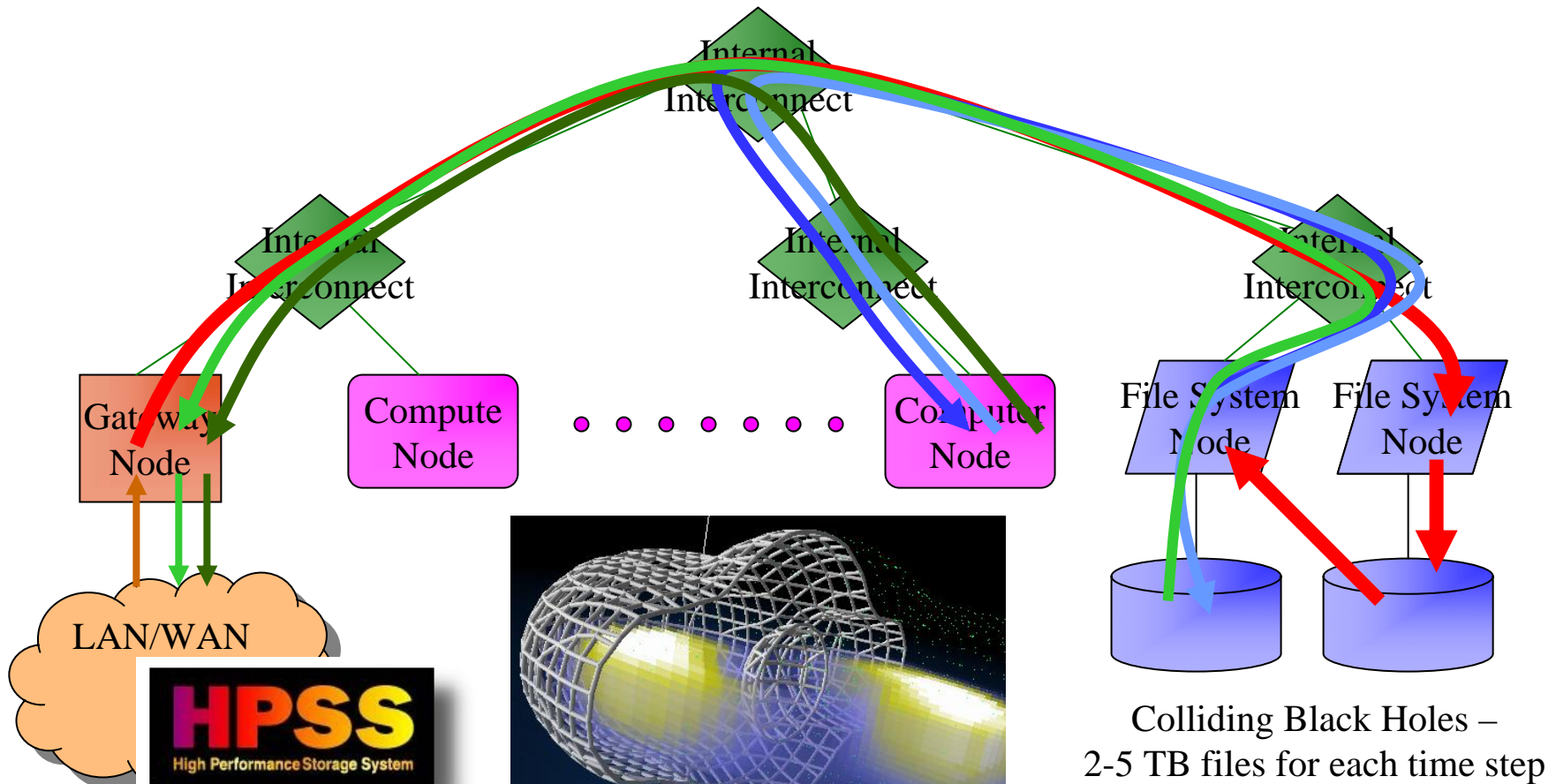
# Mass Storage Bandwidth Needs As Well





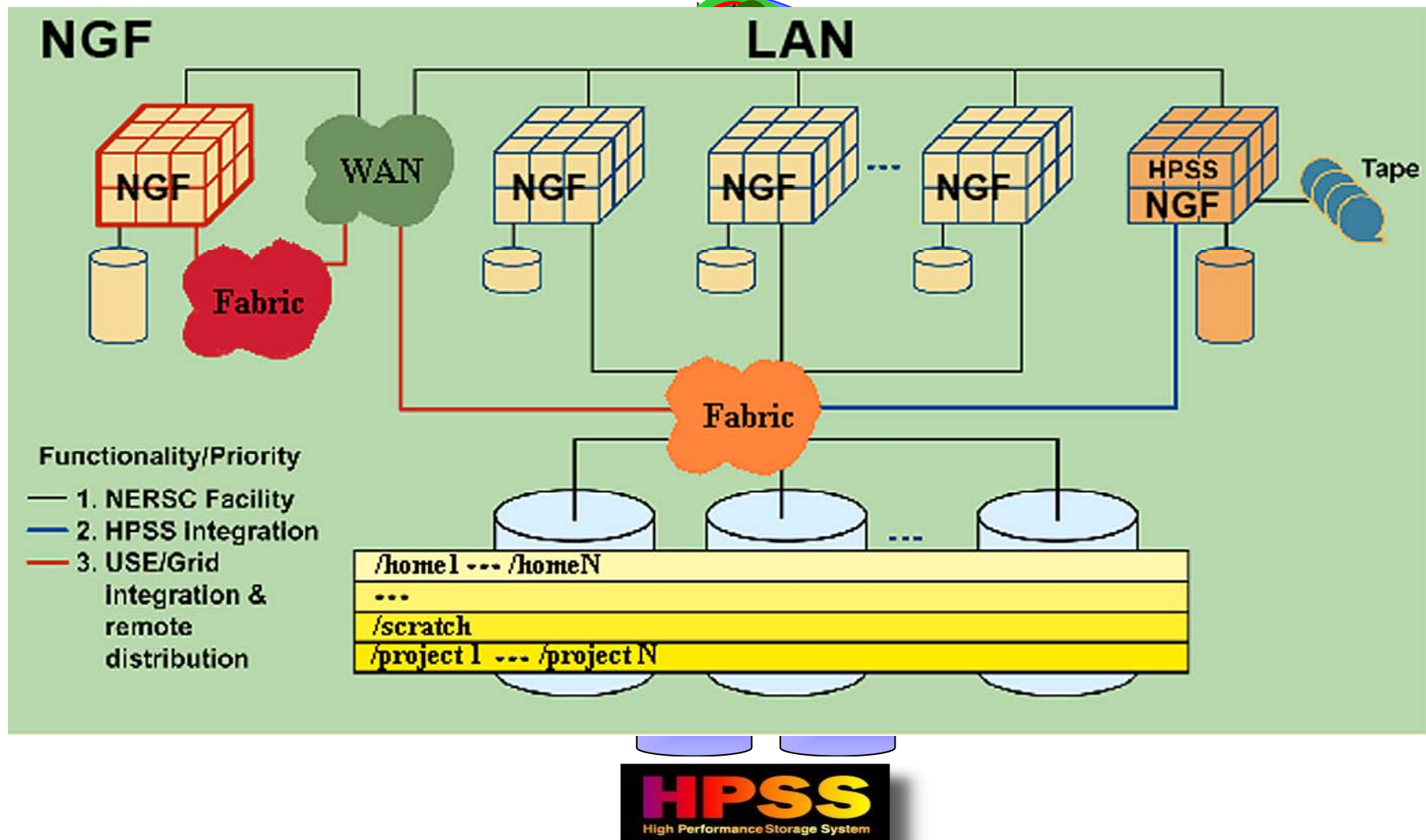
# Data Divergence Problem

The memory divergence problem is masking the data divergence problem





# NERSC Global Filesystem





# NERSC Global File System

- **A production, Center-wide, high performance, shared file system at NERSC**
  - Makes scientific research using NERSC systems more efficient and productive
  - Simplifies user data management by providing a shared disk file system in NERSC production environment
- **Global/Unified**
  - A file system shared by major NERSC production systems without replication
  - Uses consolidated storage and provides unified name space
  - Integration with HPSS and Grid is highly desired
- **Parallel**
  - File system provides performance that is scalable as the number of clients and storage devices increases at near native storage rates

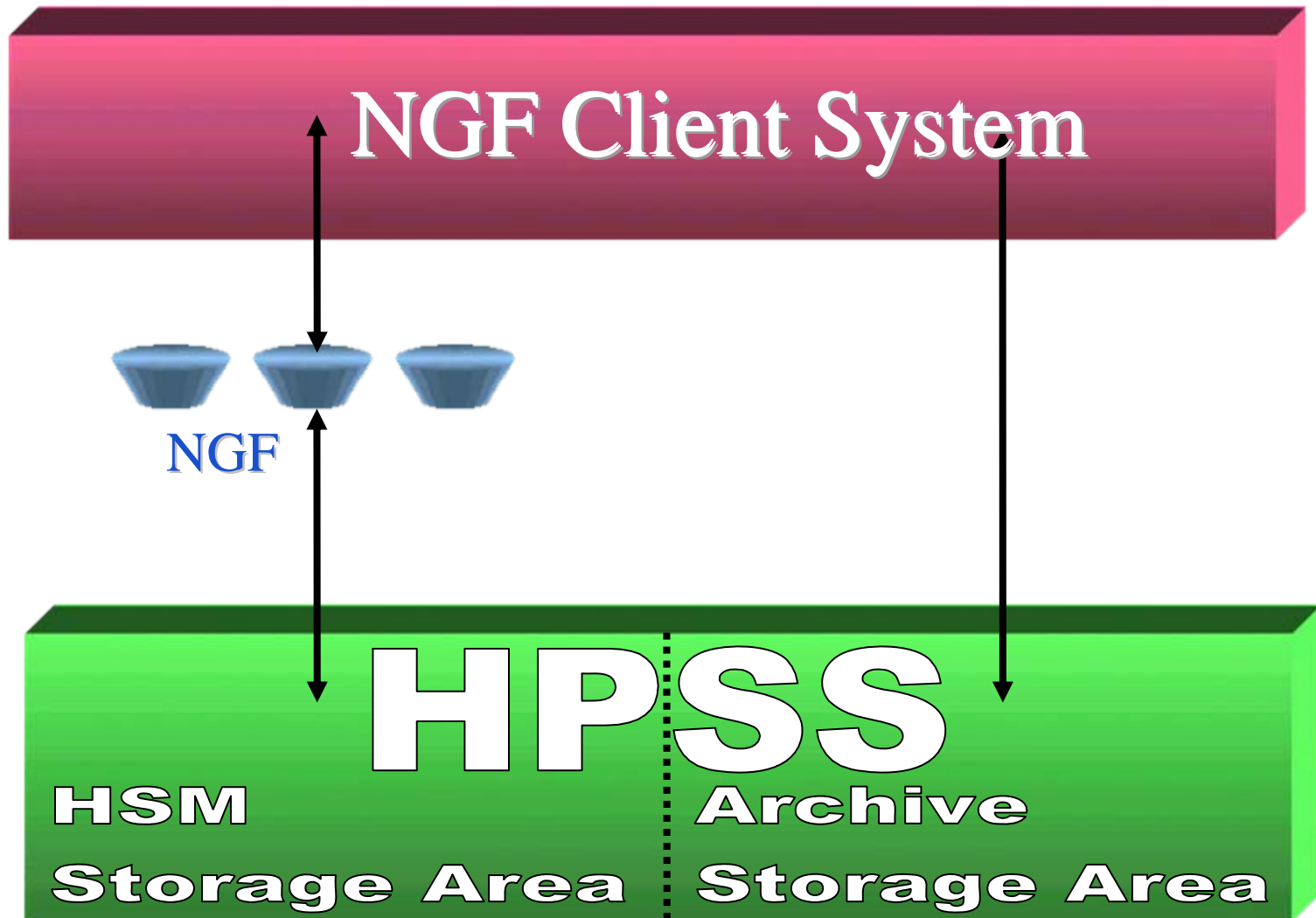


# NGF Deployment

- **FY 02-05: Investigation of storage devices, connection fabrics and storage systems.**
  - Fabrics can interoperate
  - File System software is key – especially for heterogeneous computing environments
- **Now:**
- **70 TB usable end user storage and 3 GB/s bandwidth for streaming I/O**
  - Initial clients are intended to be:
    - Seaborg, IBM Power 3+ SP running AIX 5.2
    - Jacquard, LNXI Opteron System running SLES 9
    - Da Vinci, SGI Altix running SLES 9
    - Bass, IBM Power 5 SP running AIX
    - PDSF IA32 Linux cluster running RHEL
  - Storage and servers external to all client systems
  - Distributed over a 10 Gigabit Ethernet infrastructure
  - Single file system instance providing file and data sharing among multiple client systems
    - Both large and small files expected
    - Not a scratch or a home file system
  - Focus on function first, then performance
  - Additional storage later in FY 06
- **IBM agreement to make GPFS widely available on any equipment.**
- **NERSC-5 RFP required all vendors to integrate with GPFS and all proposals did.**
  - Even some vendors who did not bid are now going to provide GPFS as part of their system.
  - Based on technology – the disk may be shared as a system-wide FS and the NGF – or all NGF.
- **Evolving to a single subsystem that has all user data transparently and equivalently available regardless of what NERSC system they are using.**



# NGF HSM and HPSS Archive





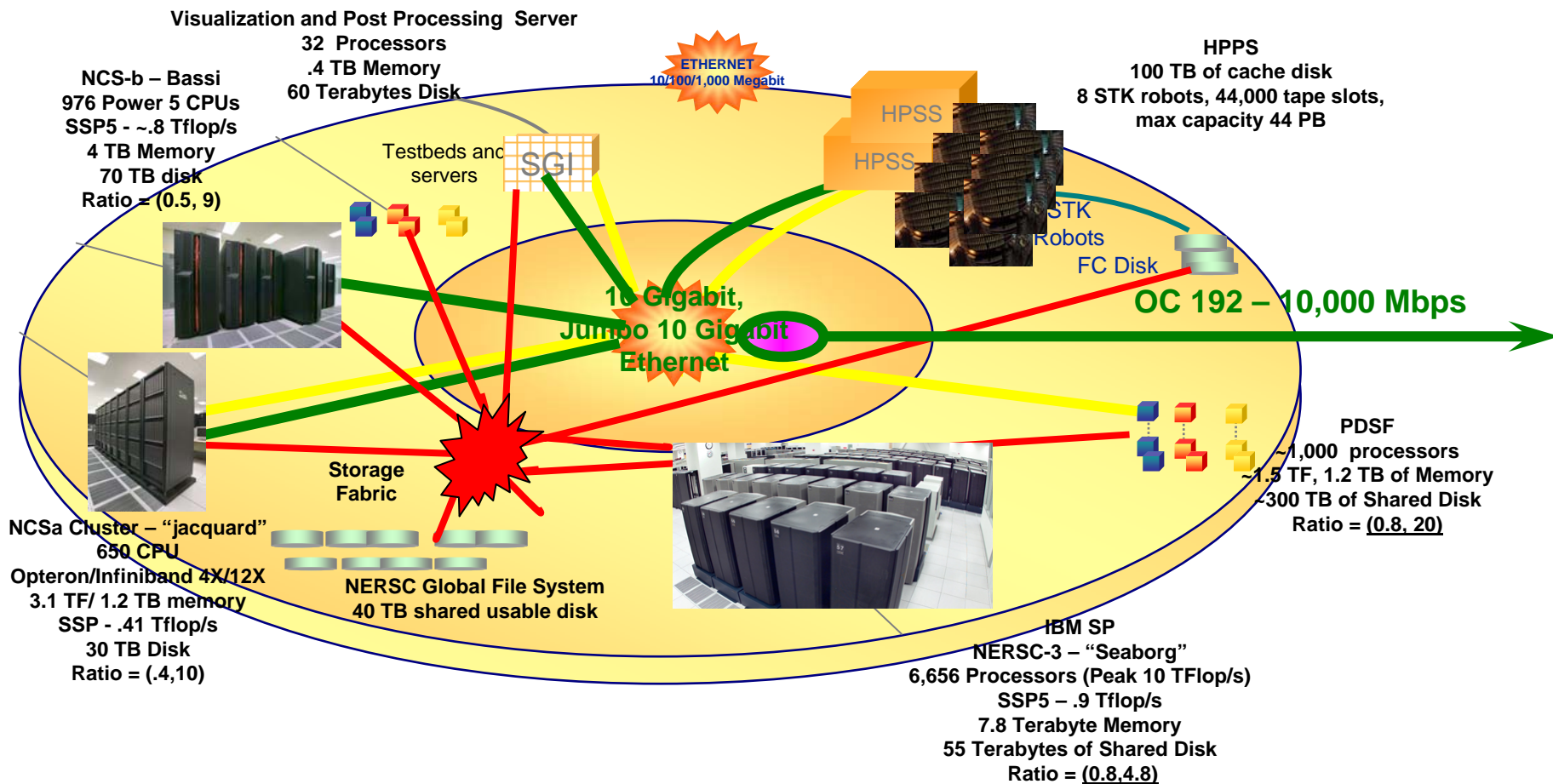
# NGF and HPSS

- **NERSC, IBM and SDSC are engaged in creating a high performance interface between NGF and HPSS.**
  - **Demonstration at SC|05 for the AIX version**
    - An AIX Core Server, 2 LINUX Movers and a single AIX GPFS node.
    - NERSC was using a GPFS file system, and IBM was using a different file system. NERSC was using a 1-way stripe COS, and IBM was using a 4-way stripe COS.
  - The demo consisted of starting with a clean GPFS file system.
  - Scripts that write 10-20 files into GPFS.
  - Viewing HPSS showed the HPSS namespace was in sync with GPFS.
  - The GPFS files showed 0 bytes in HPSS, until the files were migrated into HPSS.
  - Once the GPFS files migrated into HPSS, the HPSS file sizes had the correct size.
  - Showed
    - the file data resided in GPFS and/or HPSS.
    - that the GPFS file system space was recovered, since the data was purged from GPFS.
    - The selected a file and staged it back to GPFS and showed that the file was now in both HPSS and GPFS.
- **Linux version schedule for Q2 06.**



# NERSC Configuration

## January 2006



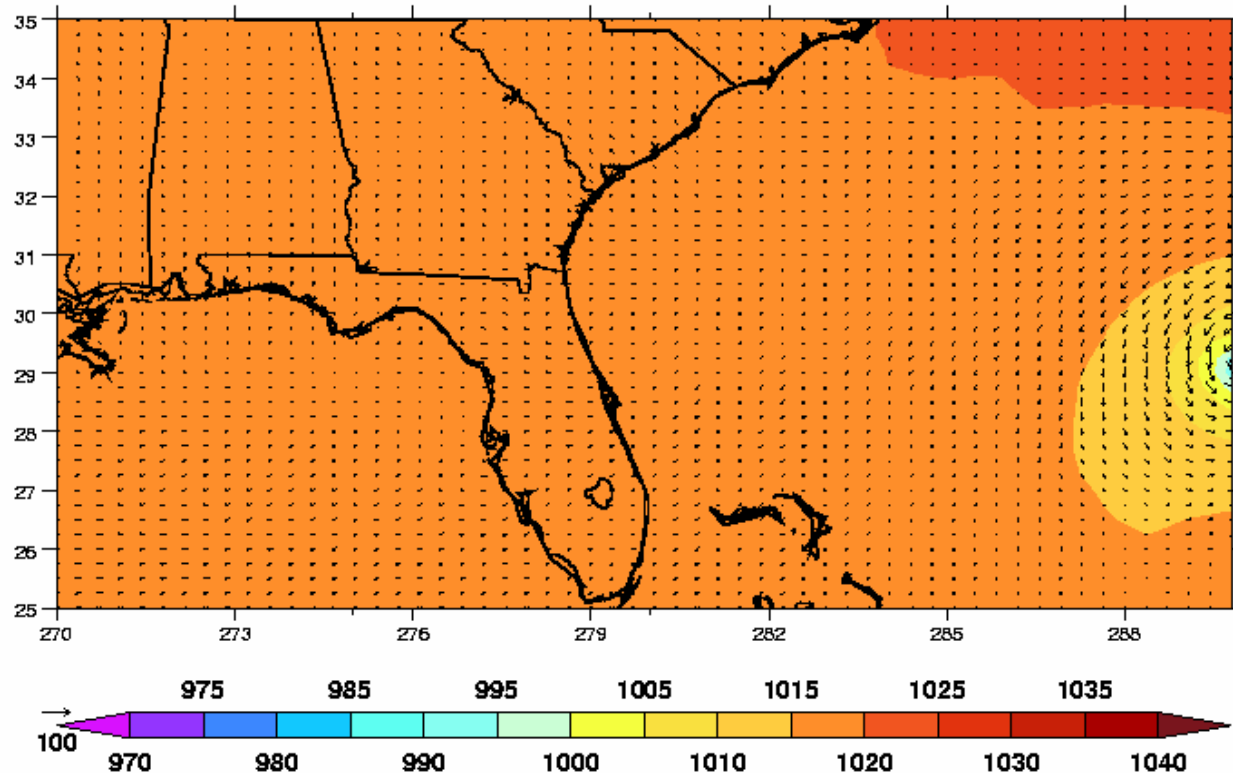
Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)



E:  $0.25^{\circ} \times 0.375^{\circ}$

Maximum surface wind speed = 84.743041587397798 mph

Minimum sea level pressure = 991.95382812499997 mb



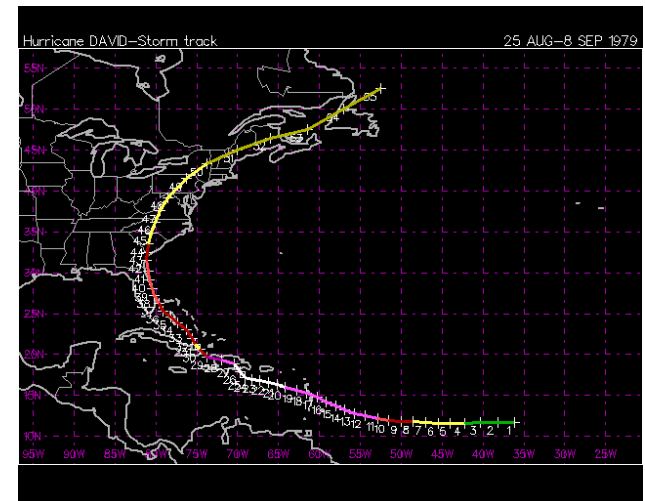
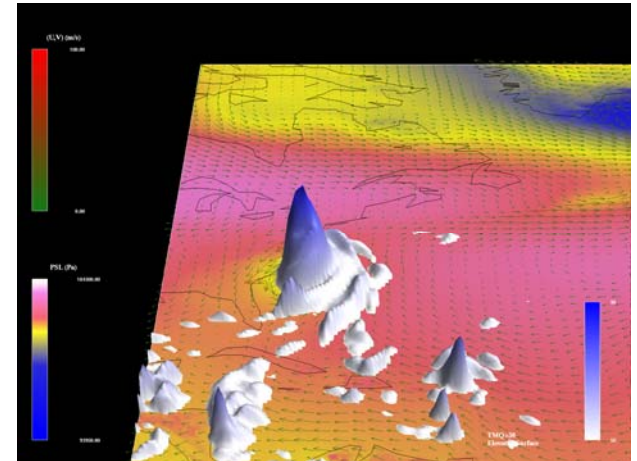
Video courtesy of Micheal Wehner (LBNL),  
Raquel Romano and Christina Siegerist (LBNL)

1979/10/2 0:0:0.0



# Comparing Real and Simulated Storm Data

- Michael Wehner (LBNL)
- The effect of climate change on the intensity and frequency of hurricanes in area is of utmost importance to policymakers.
- A workflow enabling fast qualitative comparisons between simulated storm data and real observations

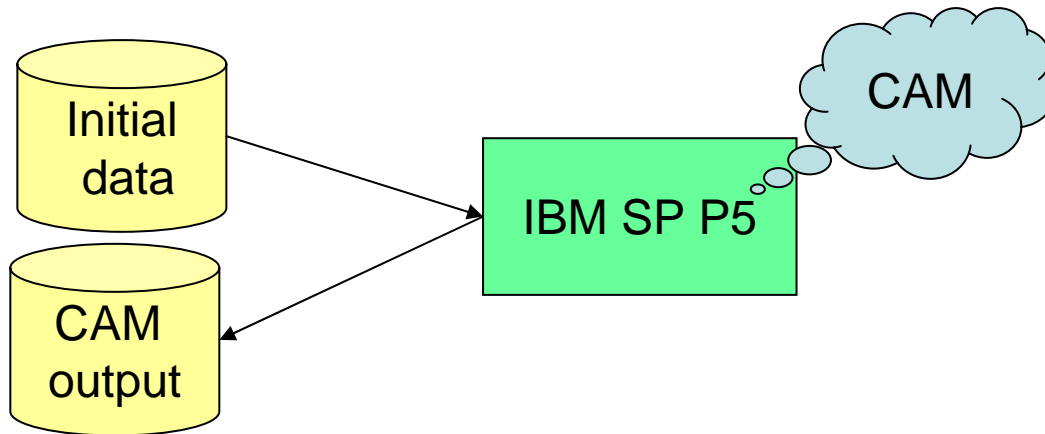




# Comparing Real and Simulated Storm Data – The Old Way

Brings most appropriate computational resource to each step of the problem in a secure and seamless manner.

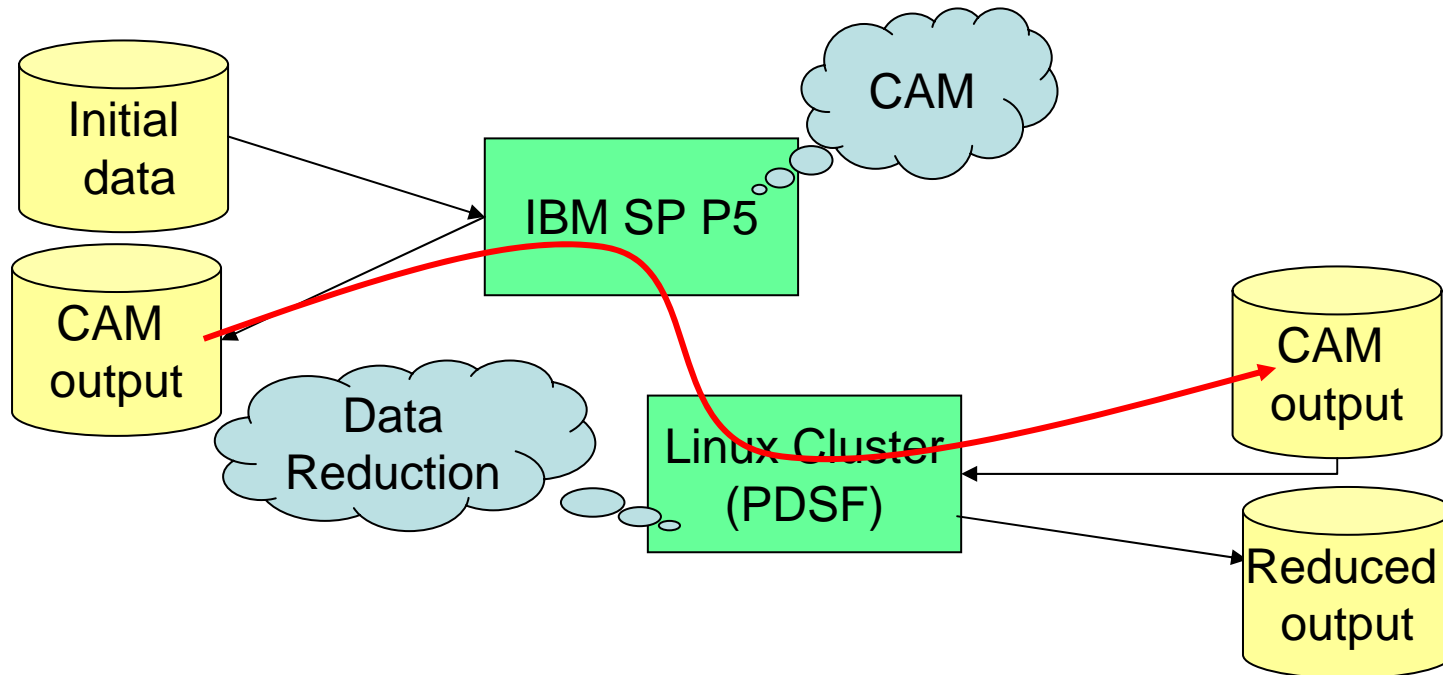
Community Atmospheric Model (CAM) runs on high performance parallel machine





# Comparing Real and Simulated Storm Data - – The Old Way

Data is transferred to commodity cluster for data reduction





```
graph LR; ID[(Initial data)] --> IBM[IBM SP P5]; CO1[(CAM output)] --> IBM; IBM --> CO2[(CAM output)]; IBM --> RO1[(Reduced output)]; CO1 -.-> IBM; CO2 -.-> RO1; RO1 --> RD[(Reduced data)]; RD --> SGI[SGI Altix]; SGI --> RD; SGI --> Display[Display]; IBM --- CAM((CAM)); PDSF((Data Reduction)) --- LCL[Linux Cluster PDSF]; SGI --- Viz((Visualization));
```

Initial data

CAM output

IBM SP P5

CAM

Linux Cluster (PDSF)

Data Reduction

Reduced output

Reduced data

SGI Altix

Visualization

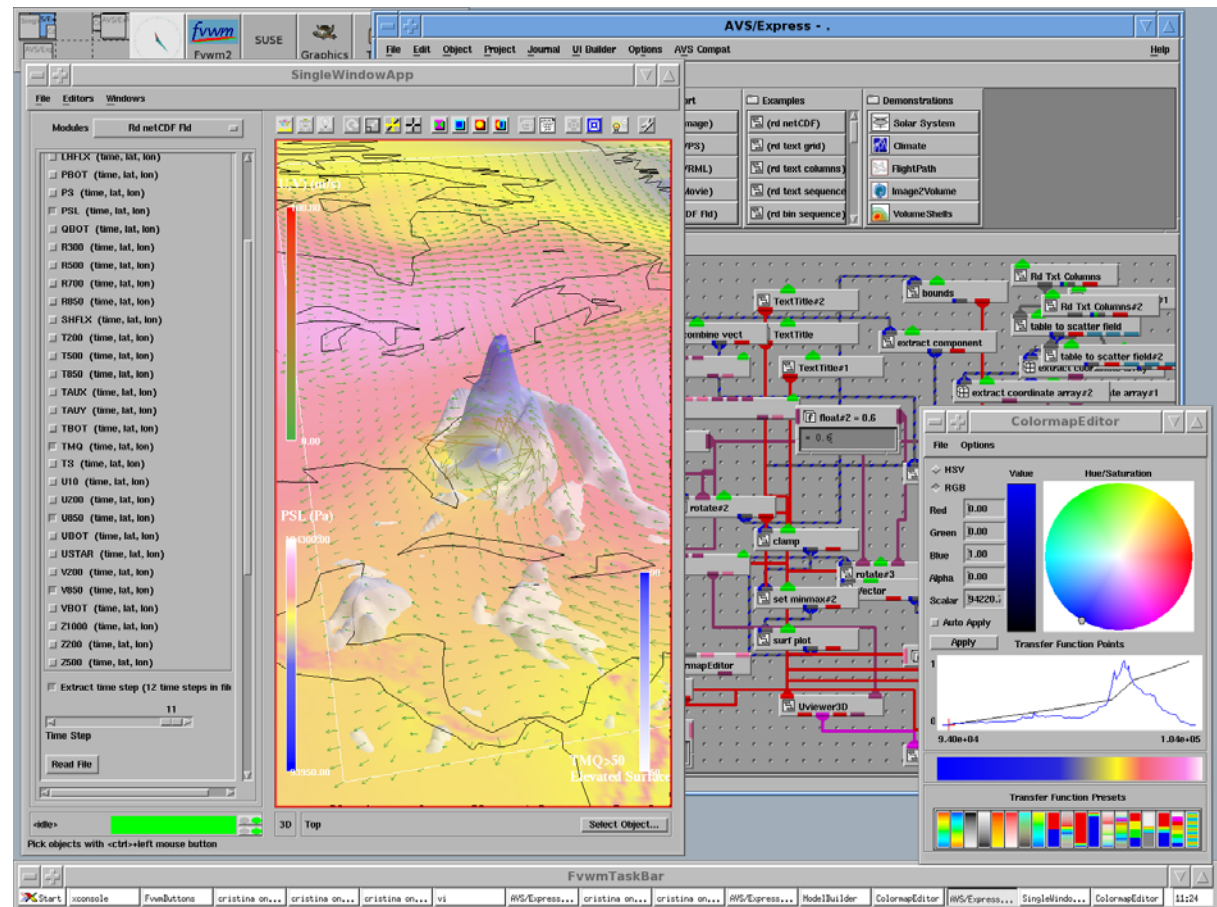
Display

Office of Science



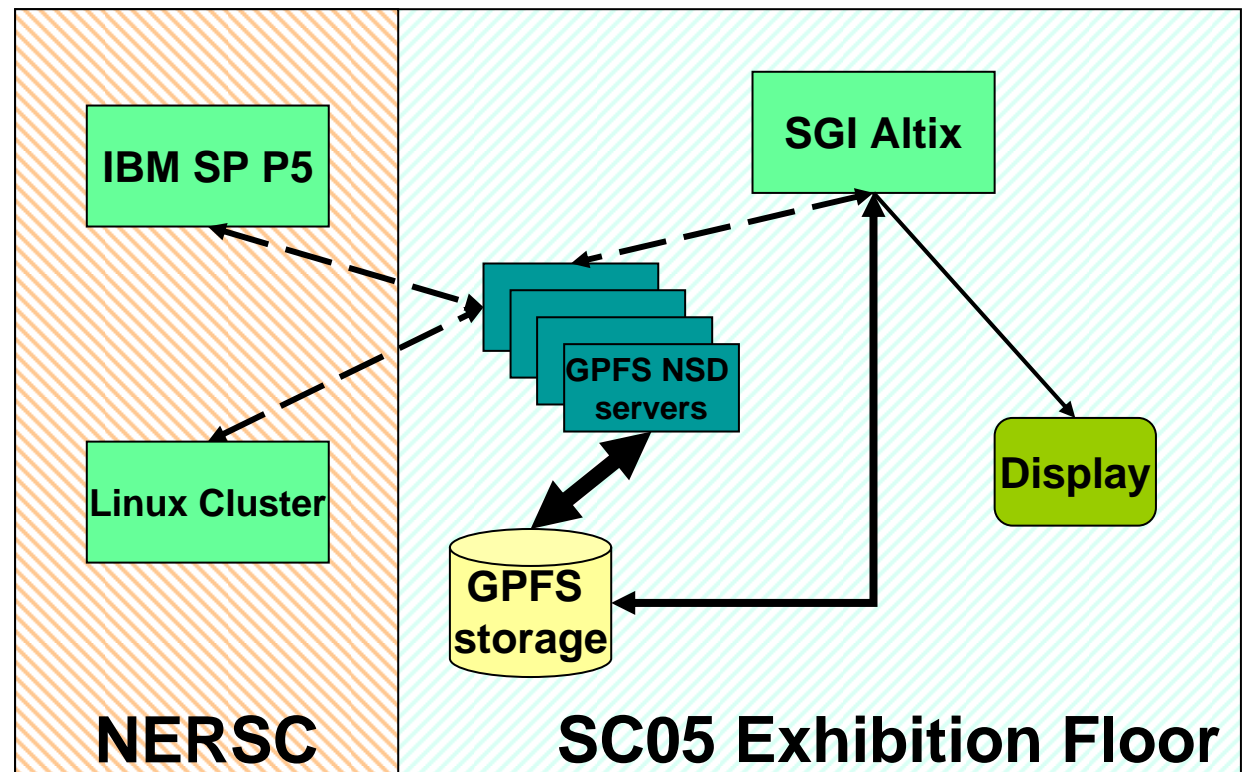
# Visualization Application

AVS/Express enables customized applications to be designed to read and display frames of climate data



# SC05 - TRI Data Storm – The New Way

- Entered prototype in SC05 StorCloud Challenge
- Separate computational resources coupled via WAN-GPFS
- Winner: Best Deployment of a Prototype for a Scientific Application  
*William P. Baird, Wes Bethel, Jonathan Carter, Cristina Siegerist, Tavia Stone, and Michael Wehner*



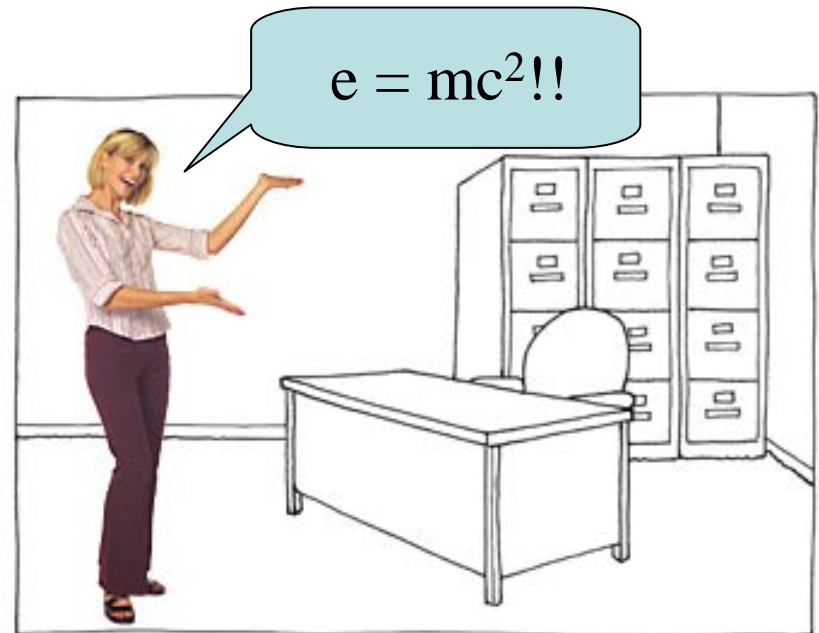


# Annual averages FV CAM AMIP 1979-1994 (except 1985)

- **Named Storms (Wind Speed >35 knots)**
  - D mesh FVCAM=11.3
  - Observed = 9.6
- **Named Storm Days**
  - D mesh FVCAM = 33.8
  - Observed = 49.1
- **Hurricanes Wind Speed >64 knots**
  - D mesh FVCAM = 2.4
  - Observed = 5.9
- **Hurricane Days**
  - D mesh FVCAM = 3.7
  - Observed = 24.5

# Conclusion

- Our Objective
  - Increase scientific productivity





# Questions and Comments